
# MINING CONSUMER REVIEWS TO GENERATE RATINGS OF DIFFERENT PRODUCT ATTRIBUTES WHILE PRODUCING FEATURE- BASED REVIEW-SUMMARY

| | |
|---|---|
| Journal: | *International Journal of Systems Science* |
| Manuscript ID: | TSYS-2014-0842 |
| Manuscript Type: | Original Paper |
| Date Submitted by the Author: | 25-Nov-2014 |
| Complete List of Authors: | Kangale, Akshay; IIT Kharagpur,<br>Kumar, SriKrishna; IIT Kharagpur,<br>Naeem, Arshad; IIT Kharagpur,<br>Williams, Mark; University of Warwick,<br>TIWARI, MANOJ; IIT Kharagpur, Industrial Engg. and Management |
| Keywords: | Bayesian Estimation < Theory-Framework, Data Mining < Artificial Intelligence |
| Keywords (author supplied): | |
| | |

# MINING CONSUMER REVIEWS TO GENERATE RATINGS OF DIFFERENT PRODUCT ATTRIBUTES WHILE PRODUCING FEATURE-BASED REVIEW-SUMMARY

Akshay Kangale
Email: akshay.kangale@gmail.com
Institution: *Department of Industrial and Systems Engineering*
*Indian Institute of Technology Kharagpur, India*

S Krishna Kumar
Email: krish5329@gmail.com
Institution: *Department of Industrial and Systems Engineering*
*Indian Institute of Technology Kharagpur, India*

Mohd Arshad Naeem
Email: arshadnaeem.iitkgp@gmail.com
Institution: *Department of Industrial and Systems Engineering*
*Indian Institute of Technology Kharagpur, India*

Mark Williams
Email: M.A.Williams.1@warwick.ac.uk
Institution: *Product Evaluation Technologies, WMG*
*University of Warwick, Coventry, United Kingdom*

M.K. Tiwari* (*Corresponding Author)
Email: mkt09@hotmail.com
Institution: *Department of Industrial and Systems Engineering*
*Indian Institute of Technology Kharagpur, India*

# MINING CONSUMER REVIEWS TO GENERATE RATINGS OF DIFFERENT PRODUCT ATTRIBUTES WHILE PRODUCING FEATURE-BASED REVIEW-SUMMARY

**Abstract:** With the massive growth of the internet, product reviews increasingly serve as an important source of information for customers to make choices online. Customers depend on these reviews to understand users' experience and manufacturers rely on this user-generated content to capture user sentiments about their product. Therefore, it is in the best interest of both customers and manufacturers to have a portal where they can read a complete comprehensive summary of these reviews in minimum time. With this in mind, we arrived at our first objective which is to generate a feature-based review-summary. Our second objective is to develop a predictive model to predict the next week's product sales based on numerical review ratings and textual features embedded in the reviews. When it comes to product features, every user has different priorities for different features. To capture this aspect of decision making, we have designed a new mechanism to generate a numerical rating for every feature of the product individually. The data has been collected from well-known commercial website for two different products. The validation of the model is carried out using a crowd-sourcing technique.

**Keywords**: Natural Language Processing; Crowd sourcing; Feature-based summarization; Opinion spam; Part of Speech Tagging; Naïve Bayes; Logistic Regression; Classification

## 1. Introduction

Online retailers provide a platform to their customers to give a review of the product or service which customers have purchased from them. With increasing use of internet, the online retailing industry is growing day by day. The way of shopping for goods is being transformed from "traditional retail" to "online retail" at very fast rate. The online retailing differs from traditional retailing in many ways. In traditional retailing, i.e. buying goods from a physical shop, the customers can check and test products before buying them. But in online store, ability to evaluate product directly is very limited. Because of this reason online customers depend on various other sources of information. These sources can be either "word of mouth publicity" or "user- generated reviews".

In addition to the text written by reviewer, there is some additional information about the product which is added by vendors. Vendors provide each and every specification of their product. But user-generated reviews attract more interest than these specifications added by vendors. In user- generated reviews, users describe the product on basis of their experience. These user-generated reviews are more user-oriented as compared to the traditional vendor information (Chen & Xie, 2008).

For new customers, user-generated data is more trustworthy and credible as compared to traditional vendor information. But there may exist thousands of reviews for popular products (Hu & Liu, 2004) and hence it is very difficult for a new customer to read all these reviews and make an informed decision.

If a customer is not reading all the reviews and buying a product after reading some reviews, he may become a victim of biased opinion of those few reviewers. Therefore, from a new potential customer point of view it is mandatory to have some kind of summary, reflecting opinions of all the reviews which have been written about the product. The manufacturers can also use these user-generated reviews to tap the latest trend and opinion of the customers about their product. Because of huge number of reviews it becomes difficult for manufacturers to read all the reviews and make management decisions for their business. So it is necessary for both customers and manufacturers to have some kind of portal where they can read summary of all the reviews in minimum time and take their individual and informed decisions. As E-Commerce is growing day by day, the amount or number of these reviews is increasing at an exponential rate. In this paper we aim to develop a system which converts these reviews into a fully furnished review-summary. This summary generation varies from general text summarization because we

*Corresponding author. Email: mkt09@hotmail.com

only extract features that are talked about by customers and their opinions. We are also interested in finding out whether these opinions are positive i.e. customer likes the particular feature or negative i.e. customer dislikes that feature. This kind of summary is known as features-based summary. It is explained step-wise as follows:

1) Identifying features about which the customers are talking about and giving their opinions.
2) For every feature, predicting whether the sentence containing that feature is positively or negatively oriented.
3) Producing a feature-based review-summary using the mined information.

With an example we will see how feature-based summary of smartphone looks like. The camera quality and battery life are the two product features among many other features as shown in Figure 1. Positive opinion about the camera quality is seen in 253 product reviews, and 6 reviews are expressing negative opinion. While in case of battery life there are 134 positive review sentences and 10 negative review sentences. For every feature of the product the proportion of positive and negative opinion is also provided separately. This proportion will be used to generate individual feature based numerical review rating. With such a summary, customer can see how much users have liked or disliked a particular feature. If they are interested in individual review sentences of particular feature, then he/she may scroll down the summary to get those sentences. In this type of feature extraction with summary generation, reviews are not summarized by selecting or writing original sentences from the reviews as it is done in traditional text summary generation.

<<< Insert Figure 1 >>>

In this paper, we have also carried out the econometric modeling of these reviews. The basic empirical equation has been taken from Archak, Ghose, and Ipeirotis (2011) but the evaluation of each feature is different in our paper. In the empirical studies such as Ogut and Onur Tas (2012), it has been proved that product reviews are considerably significant for product sales. Intuitively it can be understood that how these reviews generated by users can be significant for product sales. To capture the economic impact of reviews which are generated by users, the weight that customers put on individual evaluations and product features has been identified.

<<< Insert Figure 2 >>>

Initially online customers have certain belief about the product. Figure 2 shows how product reviews can change the person's belief about the product. It can be seen in Figure 2 that customer's belief is updated with each incoming positive product review. The consumer has initial belief of 0.5. After reading some positive reviews this belief is updated to 0.75 and thus moving towards 1. With each incoming review this belief is updated. We have considered a classification model to predict an increase or a decrease in next week's sales of a particular product. Both numerical features and textual features are taken into account to fit the model.

When it comes to features, every consumer has different preferences. For example, in case of Smartphone one customer prefers processor over camera while other customer prefers camera over processor. To capture this aspect of decision making we have designed a mechanism to generate individual numerical rating for individual feature.

Structure of rest of the paper is as follows: Detailed literature survey is presented in Section 2 while Section 3 describes the motivation behind the problem and the problem itself in detail. The data collection mechanism and preprocessing of data are explained in Section 4. The algorithms to produce feature-based review-summary are discussed in Section 5 while Section 6 explains econometric modeling of product reviews. The results have been discussed in Section 7 and finally, Section 8 concludes the paper with scope for further research.

## 2. Literature Review

In last few years the interest in mining consumer opinion has increased at an exponential rate, both in academia and industry. Very few studies have investigated about the importance of textual information hidden in user-generated reviews. Ghose, Ipeirotis, and Sundararajan (2007) study how price premiums charged by sellers in online second-hand markets are impacted by buyer textual feedback. Eliashberg, Hui, and Zhang (2007) combined NLP and machine learning techniques to predict the return on investment (ROI) of a movie. They have extracted some intuitive features from the scripts of the movie. Their model predicted the relationship between script features and success of the movies with over 90% accuracy. Netzer, Feldman, Goldenberg, and Fresko (2012) have combined text mining and graph analysis to study the association between brand network and market structure.

Mining of opinions from reviews has become a popular area of research. But, there are very few studies on review spam and web spam. Review spam is very much similar to web spam. Web spam is mainly of two types: content spam and link spam. Taxonomy of web spam is discussed in Gyongyi and Garcia-Molina (2005). E-mail spam is also a related field in which e-mails are classified into two classes: spam and normal (Fette, Sadeh, and Tomasic 2007; Sahami, Dumais, Heckerman, and Horvitz 1998). Some of the spam studies have also been extended to recommendation systems as given in Mobasher, Burke, and Sandvig (2006).

The summarization system is closely related to work of Ramezani and Feizi-Derakhshi (2014). Morinaga, Yamanishi, Tateishi, and Fukushima (2002) have compared the reviews of various products which lie in one category to come up with the image of target product. Along with it representation of opinions can also be created.

Decker and Trusov (2010) have used text mining to study the impact of product features and brand name on the total evaluation of the product. Study close to this research is done by Ghose and Ipeirotis (2011), who discuss multiple facets of review text such as semantic, lexical, grammatical and stylistic levels to find out significant text-based features and analyze their effect on product sales and review helpfulness. They have not considered the impact of these factors on the product sales. It has been seen in literature that the early preferences of first buyers can have a long term effect on late buyers (Li & Hitt, 2008). Furthermore, it has been shown that there is bimodal distribution underlying the majority of reviews. In these conditions the numerical ratings given to the reviews does not reflect completely the mood of buyer. There is additional information embedded in these reviews. As internet is growing, the importance of user- generated content has been increasing drastically. Before buying any goods, the customer always goes through reviews written about that product. These reviews show the experience of previous customers who are users of the product. It is mandatory to mine these features, so that the effect of these opinions can be analyzed properly.

The challenge is to build relation between the qualitative and quantitative part of the reviews. The qualitative part captures the textual features. It tells about the customer experience about the product usage. To solve this problem, we have to address these questions first.

1. How to identify which features are being talked about in the reviews?
2. How to identify what has been said about these features i.e. opinion of customers about the features?
3. How to capture the effect of product attributes on product sales?

With the growth of product reviews, a new area has been emerged i.e. opinion mining. For example if a marketing manager of a Nike shoe wants to know the customer opinions all over the web about its new product. This is called sentiment analysis. It can be very important for business growth of companies. The reviews can either be positive or negative (Das & Chen, 2007). Recently it has been found that, users sometimes write mixed reviews where they praise some feature of the product and criticize others. It is important to consider this aspect of reviews. Cao et al. (2010) described the “Flexible Frameworks for Actionable Knowledge Discovery” in review mining. This started additional research in finding product features in this type of content. It has been seen in literature that the early preferences of “first buyers” can have a long term effect on “late buyers”. Pang and Lee (2008) have tried to summarize all the work

that has been done in sentiment analysis. The present work in this field does not capture the impact of product reviews on product sales. But intuitively, it can be felt that product sales are dependent on product reviews up to a large extent (Liu, Hu, & Cheng, 2005).

## 3. Problem Description

Product reviews generated by users are vital source of product information for the new customers in today's era of internet. These reviews have become the basis of customer's decision for buying a particular product. The new customers read these reviews to know the experience of actual users of those products. With the help of these reviews they come to know about various pros and cons of the product. A new customer weighs the pros against cons and makes a purchase decision accordingly. Manufacturers also use these reviews to tap the opinion of customers about their products all over the web. This helps them to improve their product quality in future. Over the past decade, internet has penetrated to a very large section of society. Therefore, number of reviews is increasing at a very fast and increasing rate. For popular products, these reviews are in quantities of hundreds or even in thousands. It is tiring for both users and manufacturers to read all the reviews to facilitate their respective decisions. Hence, it is beneficial for both customers and manufacturers to have a common portal where they can get summary of all these reviews in minimum amount of time. With this motivation, we came up with our first objective which is to generate features-based review-summary in the format given in Figure 1.

During the last few years, the hypothesis that "product reviews affect product sales" has been proved by many researchers. Intuitively, this hypothesis can also be derived. Before buying any product online, customers read these reviews to make their purchase decision. If the majority of the reviews are positive, then there is a high chance of customer buying that product. Correspondingly, if the majority of the reviews are negative, there is very less chance of customer buying that product. With this motivation we come up with our second objective which is to carry out the econometric modeling of the product reviews and predict sales based on it. When it comes to features every customer is different. Some customers prefer some feature over the other while other customers prefer opposite. To capture this aspect we came up with our third objective which is to generate individual numerical ratings for individual features.

## 4. Data Collection and Pre-Processing

During the last five years, there has been huge increase in the number of online shopping websites. All these websites allow their users to write reviews (general format is given in Figure 3) about the products they have sold. In these reviews, users write about their usage experience about the product. Most of the users give opinion on different product features, whether they like it or not. They also write about the problems faced by them in the product usage. To write a review, almost similar procedure is to be followed on all the online retail websites. Before writing any review, reviewers register themselves on the site. The registration can be done in two ways which are:

- Registration on the website using user's existing social media account
- Registration on the online retail website itself

Some of the websites provide application programming interface for obtaining the reviews, but some of them do not provide any API. To get the reviews from those websites, web page scraping can be used. We have collected reviews from one of the popular website for smartphone and digital camera. We have used Beautiful Soup, a python library designed for web page scraping. Our database consists of approximately 1000 reviews for each product.

<<< Insert Figure 3 >>>

Before using the review database for computation purposes, the reviews are pre-processed first. The raw database is passed through opinion spam filter to detect the fake reviews.

Although the user-generated content is very useful, but in addition to this useful information there is vast amount of spamming. The reviews of the product play an important role in purchase decisions for new

customers which results in financial gains for the company. This gives a good incentive to write fake positive opinions about the product or to write fake negative reviews to defame the product. The spam detection is carried out in two steps as given below:

### 1) Identifying types of Review Spam

The objective of this step is to identify how many types of spam are there so that an effective system can be developed for detecting these spams. Three usual kinds of spam are discussed below.

**Type 1 (Fake Opinion Reviews)**: These types of reviews contain wrong information about the product. These are very harmful and most difficult to detect because they seem real. A very sophisticated model has to be developed for such type of reviews. They are of two types;

a. *Positive spam review:* These types of reviews reflect a false positive opinion that the product does not deserve.
b. *Negative spam review:* These types of reviews express false negative opinions. These reviews are aimed at defaming the product. These are very harmful for the product.

**Type 2 (Brand Reviews)**: This kind of reviews does not express opinion about the product. They only convey opinions about the brand of the product. This is an opinion review but opinion is on brand rather than the product. Therefore, it is considered as spam.

**Type 3 (Non-Reviews)**: This kind of reviews contains no opinion, and thus should not be considered as reviews. In general they will not affect any human reader because they are very easy to identify manually. But they can hinder the performance of automated opinion mining system which is aiming at segregating the numerical review ratings of all the reviews. These reviews are also assigned a review rating which may just be randomly given. In this type following two sub-categories exist:

a. *Advertisements:* In these reviews, the reviewers list a bundle of product features. They do not show opinion on any feature, rather they are just advertisements of the product telling about its various features. They do not have any wrong information but they are considered as spam because they do not contain opinion, therefore not serving the purpose of reviews.
b. *Other non-reviews*: Remaining non-reviews consist of following types:

*Question or Answer*: Instead of expressing opinion about product, the reviewer writes or answers a query related to the product.

*Comment*: The reviewer is commenting on some other review as a reply.

*Random Text*: This kind of review contains out of context arbitrary text completely irrelevant to the product.

### 2) Review spam analysis and detection

The two most important dimensions of a review are: reviewer information and review text. Both these data can be used to predict whether a review is spam or not. In the first case, if there is some access to the social media activities of the user, then those activities can be inspected to find out whether the user is fake or genuine. In second case, the textual content of the review is inspected by deploying various Machine Learning techniques.

Reviewer Inspection:

In this step, spam user is detected based on the online activities of the user. When a user registers through his/her social media account, then we can have access to the online activities of these users with the help of APIs provided by different social media websites.

Review Spam Detection:

In this spam detection is performed based on actual text content. It can be considered as a two class classification problem which includes classes spam and non-spam. There are various algorithms for classification problems. Naïve Bayes is best when there is text based classification. As an input, a training dataset containing all the three types of reviews, all of which are manually labeled either as spam or non-spam will be needed.

The technical details for detecting spam reviewer with the help of their social media account (Social media account provides API to get the data of user activities) is given in the subsection given below. This method can be extended to any other social media website.

### 4.1 Detecting fake reviewers

It is not easy to detect the fake users, but it is not very difficult as they all follow same patterns. The purpose of this step is to study those patterns and learn those by a suitable machine learning algorithm. The first step in any machine learning method is to extract features that can predict target variable. So, we have extracted some features which can help us to predict whether user is genuine or not. The features decided for detection of spam include three network based features and two text based features. From social networking sites we have extracted three network based features to tap the "following" relationship among users. We have extracted two text based attributes with the help of user's 20 most recent comments. In many social media sites, data can be accessed using their own Java-API.

The Java implementation of one of the popular Java-APIs is given in Figure 4. We have extracted three types of information:

1. User General Information
2. Graph/Network Based Features
3. Text Based Features

This user information will be investigated to find out whether the user is genuine or not. As explained earlier fake users always follow some kind of pattern and genuine users always depict randomness.

<<< Insert Figure 4 >>>

**Network Based Features**

To build their social network users follow their friends and allow others to follow themselves. When a user follows other user's account, he/she can get other user's updates on his/her account feed automatically. Spammers take the advantage of this feature to fetch particular user's attention by following his/her account since social media website informs that user through email about who is following whom.

Three base features have been extracted to detect fake or spam bots. These features are: number of followers, number of friends and legitimacy ratio. Some user becomes your follower if he/she follows your account to get your updates on their feed. If you are following some user's account then that user becomes your friend. For every user we have scraped both features i.e. number of followers and number of friends. The legitimacy ratio is a measure to find out the degree of truthfulness of user. The legitimacy ratio is given in Equation 1.

$$Legitimacy\ Ratio(LR) = \frac{N_{fo}}{N_{fo} + N_{fr}} \qquad (1)$$

*where*,

$N_{fo} = Number\ of\ Followers$

$N_{fr} = Number\ of\ Friends$

If the number of followers is high in comparison to the number of people you are following then the value of this ratio will tend towards 1. On the other hand, if the number of followers is very low in comparison to the number of people you are following then this ratio will tend towards zero. In the former case, when this ratio is high (Equation 2), then probability of user being fake will be very low because if the followers count is high as compared to people you follow, then it shows the popularity of the user and this in turn will increase the probability of truthfulness. In the latter case when this ratio is low (Equation 3),

then probability of user being fake will be very high because if the followers count is low as compared to people you follow, then it shows the typical behavior of spammers.

$$LR \to 1 \to P(GenuineUser) \to High \qquad (2)$$

$$LR \to 0 \to P(GenuineUser) \to Low \qquad (3)$$

**Text Based Features**

To extract text based features, we have collected twenty recent comments from the feed of every user who is being investigated. Two text based features have been mined from these twenty comments. Before mining features, we have studied timelines of fake or spam users and came up with following conclusions regarding their behavior. Firstly, the genuine users will not post duplicate their comments. But on the other hand, there are a huge number of duplicate comments from a spam account. Usually a spam account looks Figure 5.

<<< Insert Figure 5 >>>

Secondly, if some account mainly consists of HTTP links rather than updates then that account is considered as spam. To count the number of links, we have determined how many of the twenty comments contain HTTP links. If an update contains "http://" or "www." then that update is considered having HTTP link. The two text based features are: number of same or identical comments and number of comments containing HTTP link. To find out identical updates we have used Levenshtein distance. According to Levenshtein distance definition, it is a string metric for calculating the difference between two sequences.

The Levenshtein distance between two words is the minimum number of single-character edits that is insertions, deletions or substitutions which are required to transform one word into another. Levenshtein distance is also called Edit distance. Mathematically it is defined as Equation 4. The distance becomes zero only when two comments are identical. We have calculated this distance to find whether the comments are similar or not. In spam social media accounts, the presence of these duplicate comments is almost sure. The Equation 4 shows how Levenshtein distance between strings 'a' and 'b' can be calculated.

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & if\ \min(i,j) = 0 \\ \min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{a_i \neq b_j} \end{cases} & otherwise \end{cases} \qquad (4)$$

In the above equation, first element in the minimum corresponds to deletion (from 'a' to 'b'), the second to insertion and the third to match or mismatch, depending on whether the respective symbols are the same.

**Total attributes for spam reviewer detection**

Overall, we have mined three network based and two text based attributes which are:

(1) Number of Followers
(2) Number of Friends
(3) Legitimacy Ratio (LR)
(4) Number of identical tweets

(5) Number of tweets containing HTTP links

These five features will act as input to different machine learning algorithms to classify users into genuine and fake. We have applied Naïve Bayes algorithm for classification. Naïve Bayes shows best performance for these types of classification problems (He, Zhang, Li, & Shi, 2012). As explained in this section and previous section we are investigating both review content and reviewer's social media account to detect spamming. If for some review both review and reviewer turn out to be spam, then that review is removed from database. If anyone among review content and reviewer turns out to be spam, then in this condition also review is removed from database. There are some reviews for which there is no link to their social media accounts. For such users only review content is investigated.

## 4.2 Detecting Fake Reviews

In this section, a methodology is developed to classify reviews into spam and genuine by inspecting the actual text content of the review. We have used Naïve Bayes algorithm for this purpose. A corpus of manually labeled reviews is taken as training data to learn the patterns and to differentiate between spam and genuine reviews.

The Naïve Bayes classifier applies to learning tasks where each instance x is described by a conjunction of attribute values and where the target function can take on any value. A set of training examples of the target function is provided and a new instance is presented, described by the tuple of attribute values:

$$< a_1, a_2, .....a_n >$$

Target value is predicted for this new instance by the learner. The Bayesian method is to classify the new data by assigning the most probable target value provided its attributes which describe the instance. In case of document classification, the attributes are the important words present in the document. The assumption of Naïve Bayes is that the words position does not have any effect on the target value. To find these important words or concepts a pre- processing of document or review is performed. First the stop-words are removed and then every remaining word is stemmed. These remaining words are the concepts of a document. Let these words be represented by:

$$< X_1, X_2, .....X_n >$$

According to Bayes theorem, the probability of a review being spam given its important words, can be written in the form of Equation 5.

$$P(spam \mid X_1, X_2 ... X_n) = \frac{P(X_1, X_2 ... X_n \mid spam) * P(spam)}{P(X_1, X_2 ... X_n)} \qquad (5)$$

Where $P(spam \mid X_1, X_2 ... X_n)$ is called as posterior probability and $P(X_1, X_2 ... X_n \mid spam)$ is probability of observing specific set of words provided that review is spam. This is calculated from training dataset. Since we have assumed that words are independent of each other and position does not determine the target value, we can state the following;

$$P(X_1, X_2 ... X_n \mid spam) = P(X_1 \mid spam) * P(X_2 \mid spam) \\ .....* P(X_n \mid spam) \qquad (6)$$

$$P(X_n \mid spam) = \frac{Number\ of\ times\ X_n\ has\ occured\ in\ false\ reviews}{Total\ number\ of\ words\ in\ false\ reviews} \qquad (7)$$

In Equation 7 if numerator is zero i.e. a particular word doesn't occur in the review then a small value is taken in numerator rather than zero. Equation 8 gives the probability of spam.

$$P(spam) = \frac{Number\ of\ \ false\ reviews\ in\ database}{Total\ number\ of\ reviews} \quad (8)$$

In this application, the total number of false reviews and truthful reviews is taken as equal. As explained above, the probability of a review being genuine given its attributes, can be written as:

$$P(normal \mid X_1, X_2 ... X_n) = \frac{P(X_1, X_2 ... X_n \mid normal) * P(normal)}{P(X_1, X_2 ... X_n)} \quad (9)$$

The denominator is common in both Equation 1 and 5. So, the left hand side posterior probabilities are directly proportional to the numerator as given in Equation 10 and 11.

$$P(spam \mid X_1, X_2 ... X_n)] \alpha [P(X_1, X_2 ... X_n \mid spam) * P(spam) \quad (10)$$

$$P(normal \mid X_1, X_2 ... X_n)] \alpha [P(X_1, X_2 ... X_n \mid normal) * P(normal) \quad (11)$$

## 5. Feature-based Review-summary

In this section we have provided each and every technical detail pertaining to the summary generation. Architectural overview of our opinion summarization system can be seen from Figure 6. Feature extraction and opinion direction identification are two main steps involved in summarization done by the system. The inputs to the system are product name and an entry page for all the reviews of the product. The output of the system is the summary of reviews which is shown in Figure 1. Before any further processing, the data is passed through opinion spam filter, which can detect fake reviews. Every review is then divided into its individual sentences. Here it is assumed that a customer expresses his/her complete opinion in one sentence. The step wise procedure for generating feature-based review-summary is shown in subsequent sub sections.

<<< Insert Figure 6 >>>

### 5.1 Part of Speech Tagging

Consider the two examples given below:
"The quad-core processor is really fast and smooth"
"The camera is really amazing. The picture quality is crystal clear"
In the first example, the reviewer is satisfied with the speed of the processor. The "processor" is the feature about which reviewer is talking about and "fast" and "smooth" are opinion words. Similarly in the second example the reviewer is talking about the camera. The "camera" and "picture" are the features while "amazing" and "crystal clear" are opinion words. It can be seen that the features are generally noun and noun phrases. Therefore the part of speech tagging is used to find out these nouns/noun phrases.
The objective of POS tagging is to determine in a simple way the grammatical function of the word. It is most fundamental and basic task used in Natural Language Processing model. Sometimes it can be hard to know the part of speech of word because there are many words which can be used as multiple parts of speech. To resolve this ambiguity, context of the word is to be identified which in turn also is a difficult task to perform. The Tag Set as given by the Penn Treebank is given in Table 1.

<<< Insert Table 1 >>>

We have used a simple Charniak model for Part of speech tagging. It gave an accuracy of 95.6% on our dataset. Probabilistic taggers such as Charniak model are used where a tagged corpus is used to train some sort of model. POS Tagging process is shown in Figure 7.

<<< Insert Figure 7 >>>

A simple Charniak model for POS tagging is shown in Figure 8.

<<< Insert Figure 8 >>>

This model calculates the probability of a word belonging to a specific tag. This probability is calculated for every word and each word with every tag and is explained in Equation 12, 13 and 14.

$$P(t^i \mid w^j) = \lambda_1(w^j)\frac{C(t^j, w^j)}{C(w^j)} + \lambda_2(w^j)\frac{C_n(t^i)}{C_n()} \qquad (12)$$

*where*

| | |
|---|---|
| $C(t^j, w^j)$ | *number of times word j appears with tag i* |
| $C(w^j)$ | *number of times word j appears* |
| $C_n(t^i)$ | *number of times a word that had never been seen with tag i gets tag i* |
| $C_n()$ | *number of such occurrences in total* |

$$\lambda_1(w^j) = \begin{cases} 1 & \text{if } C(w^j) \geq 1 \\ 0 & \text{otherwise} \end{cases} \qquad (13)$$

$$\lambda_2(w^j) = \begin{cases} 1 & \text{if } C_n() \geq 1 \\ 0 & \text{otherwise} \end{cases} \qquad (14)$$

All the words with noun tags are seen as potential features. But all are not equally important. Some of the features are "frequent" i.e. those are most talked about while some of them are moderately frequent and remaining are less frequent. To extract these frequent features or attributes we have used Association Rule Mining.

## 5.2 Frequent Feature Extraction

All the noun or noun phrases are not of interest. In this section we have illustrated how association rule mining can be used to find the "hot" features or the features that are most talked about. Let us explain the basics of association rule mining.

It aims at mining interesting correlations, frequent patterns, association or casual structures among sets of items in the transaction databases. It expresses how objects are related with other objects and how they need to be in group together. In association rule mining the data should be either transactional or relational. Each transaction has an identifier and a set of items as shown in Table 2.

<<< Insert Table 2 >>>

A simple example of association rule is:

- Beer → Snacks (0.5%, 60%)
- If the customer buys beer he buys snacks in 60% of the instances. Beer and snacks are bought together in 0.5% of the rows in the database.

"Support" and "Confidence" are two terms which are used in association rule mining and are explained in Figure 9.

<<< Insert Figure 9 >>>

Association rule mining consists of following two steps:

1) Finding the frequent item sets i.e. the sets of items that have at least the minimum support.
2) Deploying these frequent sets to find out the strong association rules that satisfy the minimum support and minimum confidence.

We have to search only for those frequent item sets which have 3 or less words because it is assumed that a feature will not contain more than three words. The feature sets obtained from association rule mining is saved in features list before further processing.

### 5.3 Feature Pruning

The frequent feature set obtained from association rule mining is passed through further pruning because not all of them are relevant. Feature trimming is used to trim some of the unwanted features. We will describe three kinds of pruning as given in Hu and Liu (2004). But we have added one extra type of pruning called Miscellaneous Pruning which further improves our algorithms.

**Compactness Pruning:** This type of pruning or trimming is used to check the features that contain at least two words, which we call feature phrases, and remove which are most probably irrelevant. The necessity of this pruning comes from the fact that association rule mining algorithm Apriori does not take into account the position of a word in the sentence. However words that come together in the sentence have more chance to form meaningful phrases. That's why some of the feature sets generated by Apriori may be meaningless and should be removed or pruned. The idea of compactness pruning is to trim or remove those feature phrases from the feature list whose words do not appear together.

Let us define compact phrase. Suppose *f* is a frequent feature phrase and *f* contains n words. Assume that a sentence *s* contains *f* and the order of the words in *f* that appear in *s* is: $w_1$, $w_2$,…,$w_n$. If the distance between two words in *s* that appears in *f* is not greater than 3, then we can conclude that *f* is compact in *s*. If *f* occurs in m sentences of the review and it is compact in at least 2 of the m sentences, then we call *f* a compact phrase. The non-compact phrases are to be immediately removed from the features list.

**Redundancy Pruning:** The purpose of this step is to prune redundant features which have single words. There are single word features which doesn't have any meaning when they are used alone. For example, in case of Smartphone "life" has no particular meaning when used alone but has absolute importance when used as 'battery life" or "camera life". The objective of this step is to remove such type of features.

**Miscellaneous Pruning:** In this pruning we prune those features which are similar in meaning. Only one of them is retained for further analysis. For example, "sound' and "audio" have almost same meaning. So one of them is removed and other one is retained with frequency of both. This pruning is application dependent.

After feature pruning, the features list is updated. These features are the most talked about features in the reviews. It shows that many people are interested to know about these features. That's why summary generation is really important in this domain, so those new customers use these summaries and decide on their purchase design while saving time on reading many of them. The next step is to extract opinion words.

### 5.4 Extracting Opinion Words

In product reviews, the customers express their sentiments using opinion words on different product features. The examples of opinion words are "good", "amazing", "poor" etc. These opinion words are generally located near to the feature. For instance let us see two examples:

"The display is horrible; this was not expected from smartphone manufacturer."
"The camera is awesome. It takes amazing pictures in both day and night."

In first example, “display” is the feature and it is nearby the opinion word “horrible”. In the second sentence, features “camera” and “picture” are close to opinion words “awesome” and “amazing”. To find out the opinion word nearby adjective is found out using POS tagger. If there is word “not” before that opinion then “not” will reverse its orientation. The pseudo code for opinion word extraction is given in Figure 10.

<<< Insert Figure 10 >>>

When there is one or more than one feature and opinion words in the review sentence, then that sentence can be referred as opinion sentence. The next step is to find out the semantic orientation of opinion sentence, which in turn can be find out using semantic orientation of individual opinion words.

### 5.5 Orientation of Opinion Sentences

This part presents the most important and crucial objective of this paper i.e. to find semantic orientation of the opinion sentences. It can be identified by using two steps. Firstly, the orientation of every word is found out. Secondly, the semantic orientation of sentence is calculated by combining orientation of all words.

For every word, it is essential to find its sentiment or orientation, which will be further used to predict the sentiment of all sentences. Semantic orientation of a word can be defined as the direction in which it deviates from the usual norm. Words like “amazing”, “awesome” reflect the desirable state of particular feature. These words have positive orientation while words like “poor”, “disappointing” express the undesirable state of the particular product. These words have negative orientation. Unfortunately dictionaries do not include sentimental or semantic orientation of each word. In this paper we have followed one of the famous online databases for reference, which will be called as “Database” hereafter in this paper for our convenience. We have used Database-based algorithm to predict the orientation of every opinion word. This method proves to be very effective. There are two types of sets in Database viz. “synonym set” and “antonym set”. All the adjectives, or in our case opinion words which are included in the Database are divided into bipolar clusters which can be seen in Figure 11. It consists of two separate clusters, one for synonym of fast and one for synonym of slow. Synonyms and antonyms are connected by dashed line while synonyms are connected by full solid lines. Each cluster is lead by head synset “fast” and its antonym “slow”. Satellite synset follows these head synset and represent words which have similar meaning as that of head adjective.

<<< Insert Figure 11 >>>

Usually adjectives have the same semantic orientation as their synonym and opposite orientation of their antonym. We have used this knowledge to find the semantic orientation of opinion word. To accomplish this, the synonym and antonym set of the given adjective are searched. If a synonym or antonym of an adjective has known orientation then we can predict the orientation of the adjective. This method needs a corpus of seed adjectives whose orientation has been manually labeled. If we have enough seed adjectives in the corpus we can predict the orientation of almost all adjectives. We have used a seed list of 60 adjectives with manually labeled orientation. Suppose the semantic orientation of a word needs to be found out. First step would be to find out the synonym, or antonym or exactly same word. This can be done using Database. If the orientation of antonym of any adjective is known, then the orientation of adjective will be reversed and if the orientation of synonym is known, then the orientation of the adjective is same as its synonym. When the orientation of the word is known it is appended to the seed list. So, seed list grows itself by every incoming adjective with known orientation.

We have now reached the step of finding the orientation of sentence. The orientation will be either positive or negative. Here we have used the orientation which is dominant in the sentence to be the resultant orientation for that sentence. For example, if there are more positive opinions than negative in a sentence, then the sentence is of positive orientation. Similarly if negative opinions are more than positive

opinions, then the sentence will be of negative orientation. If both are equal, then sentence can be neutral or that sentence is given orientation of the last sentence because people tend to express their sentiments in continuity. Sometimes it can be beneficial to tag the sentence as neutral, because if we are going too extreme in the prediction of orientation then the product may lose its value. The purpose of predicting sentence orientation is to use the sentence to generate feature-based summary as given in Figure 1. Note that when we find orientation of the word we simply do not take the sentimental orientation of the adjective as its orientation in the sentence. We take into account the presence of negation words like "no", "not" because these words can reverse the orientation of that opinion word. In examples like "camera is not good", the presence of "not" totally changes the orientation of "good".

## 6. Econometric Modeling and Individual Ratings

This section deals with various techniques used to evaluate the effect of customer opinions on product sales. In the dataset, we have sales data and reviews of the product. We do not have actual number of units sold for a particular product but we have sales rank. The literature shows that it can be used in place of demand. The effect of product reviews on product sales can be modeled by including feature-based review information in linear equation of sales rank. The empirical model is given in Equation 15.

$$\log(s_{jt}) = d_j + \gamma_p p_{jt} + X_{jt}\beta_{jt}^{x} + \\ + Y_{jt}\beta_{jt}^{y} + Z_{jt}\beta_{jt}^{z} + \theta \log(s_{jt-1}) + \in_{jt'} \qquad (15)$$

*where*

$s_{jt}$ *sales rank for product j at time t*

$d_j$ *product specific fixed effect*

$p_{jt}$ *price for product j at time t*

$X_{jt}$ *vector of numeric review variables*

$Y_{jt}$ *vector of textual review variables*

$Z_{jt}$ *vector of control variables*

Equation 15 contains review variables for only those reviews which were published at least one day before "t". We are incorporating one period lag in our model. The reason is that the website requires some amount of time to update the statistics of sales and therefore the effect of product reviews is not captured by current sales rank. The variable X is a vector containing numeric review variables. The following variables are included in X:

(1) Average Review Rating
(2) Total Number of reviews
(3) Total length of reviews
(4) The number of one star and five star reviews
(5) Standard deviation of review ratings.

In Equation 15, the variable is the representative for the customer's belief about the product based on product reviews. The change in Y with new reviews shows the shift in belief about the product. Each component in Y denotes a possible opinion phrase. The opinion phrase is the combination of an evaluation e (opinion words) and a product feature f (feature/feature phrase). Suppose F is a set of all

product features and E be a set of all opinion words, then the number of components in Y will be |F| x |E|. The model after incorporating textual information is given in Equation 16.

$$\log(s_{jt}) = d_j + \gamma_p p_{jt} + X_{jt}\beta_{jt}^{x} + \sum_{f\in F}\sum_{e\in E} Y_{jt}(f,e)Score(f,e) + Z_{jt}\beta_{jt}^{z} + \theta\log(s_{jt-1}) + \in_{jt}, \quad (16)$$

*where*

$Y_{jt}(f,e)$ *represents a component corresponding to the pair of feature f and evaluation e*

$Score(f,e)$ *represent the corresponding slope in* $\beta_{zt}^{y}$

We have defined a vector space for the textual review variables. The value of each component in the vector is the number of times that opinion phrase is occurring in the review text divided by the number of times the particular feature was evaluated in the review. The value is given in Equation 17.

$$Y_{jt}(f,e) = \frac{N(f,e)}{s + \sum_{\hat{e}\in E} N(f,\hat{e})} \quad (17)$$

To understand Equation 17 let us take following example:
"The phone is of high quality. The processor is fast, smooth and fantastic"
This review is represented by the components of the review space of consumer with following weights (assuming s=0):

$$Y_{jt}(quality, high) = \frac{1}{1} = 1.0$$

$$Y_{jt}(use, easy) = \frac{1}{1} = 1.0$$

$$Y_{jt}(processor, fast) = \frac{1}{3} = 0.333$$

$$Y_{jt}(processor, smooth) = \frac{1}{3} = 0.333$$

$$Y_{jt}(processor, fantastic) = \frac{1}{3} = 0.333$$

To estimate the weights in Equation 20, we have applied Generalized Method of Moments (GMM).
In econometric modeling, we have carried out the predictive modeling of sales rank. In this part the forecasting of product sales has been carried out as a classification problem. The modeling can be done as:

Classification Model

The purpose of any classification model is to predict the input variables instance into two or multiple classes. Here, only two classes have been considered for predictive modeling. The objective is to predict the sign of expression Sales Rank (t+7) – Sales Rank (t). The objective is to predict whether the sales rank will increase or decrease in the upcoming week. Logistic regression is used to solve this model.
The final objective of this study is to give individual review ratings to every product feature. The customer can use these individual review ratings to make his/her purchase decisions. The overall summary format with individual ratings is given in Figure 12.

<<< Insert Figure 12 >>>

The individual rating is generated only for those features whose support is at least five percent. For remaining features only proportion of positive and negative sentences with review sentences is provided. The individual ratings can be calculated as:

$$Positive\ Ratio_i = \frac{Number\ of\ positive\ sentences}{Total\ number\ of\ sentences} \quad (18)$$

$$Feature\ Rating_i = Maximum\ scale * Positive\ Ratio_i \quad (19)$$

First, the proportion of positive review sentences is calculated for a feature. Then this ratio is multiplied by maximum scale to obtain feature rating.

## 7. Results and Discussion

We have collected 1000 reviews each for smartphone and digital camera. Because of the space constraint we are not able to show the results of all of these reviews. Therefore, in this section we have illustrated the method on a small dataset consisting of 10 reviews of smartphone. The dataset is given as follows:

1) Man this is a master piece, so fast and smooth response to every touch on the screen, sleek design comfortable and practical size light in weight, good battery life.
2) Superb display, Graphics & Processor is fast. With latest update the images are much better especially at night see the difference with HDR image and non HDR.
3) Some product features can be described as;
   Processor === very good.
   Graphics === very good.
   Display === Good.
   Looks === Very good.
   Web browsing === Great n best.
   Camera === good in day light and average in low light.
   Front camera === average.
   Battery life === good
   Users === high end and gamers and those who don't like lags.
   Recommended for high end application and gaming users.
4) Pros and cons of this product are;
   PROS:
   Good display
   Blazing fast performance
   Decent camera, in the HDR+mode
   Pricing
   Always the latest updates.
   CONS:
   Build quality average in white variant
   No micro SD expansion
   Average battery life
   Loudspeaker is bad.
   Very light weight and camera is good.
5) Some product features can be described as;
   Looks and Handling - Excellent
   Processing speed for games and apps - Excellent
   Camera - Good

Battery backup - Average (for this screen size)
Plus latest updates.

6) Pros and cons of this product are;
PROS:
1. beautiful design
2. amazing Full HD display
3. delightful UI and fast and zippy OS
CONS:
1. ultra fragile screen
2. poor battery life
3. SMS is poorly implemented
4. bad speaker
5. patchy sound from mic
6. bad sound on audio / video recordings

7) Some product features can be described as;
Processing - Awesome
Smooth - Awesome
Graphics - Stunning
No lag no matter whatever you open.
It is butter packed in hardware.

8) Light weight, minimalistic design, awesome hardware, value for money, beats almost any other phone in the same segment and totally worth buying. Hoping to get better updates as we go ahead and mind it it's not a pc so keep your blunt judgments to yourself.

9) Drawbacks:
Device not connecting Wi-Fi.
Audio sometimes is very bad.
Battery power.
Camera is average.
Advantages:
Very good look and stylish.
Touch is good.
Light weight.
Any other version could be better option.

10) Delivery guys were good, right from the time I ordered, I could check the status of the order online and the phone is nice, with a sleek simple design, OS is really fast the best phone I have had in years.

The first step is to pass this dataset through opinion spam filter. Every review is classified as genuine in the filter. The next step is to transform these reviews into individual sentences. For every sentence, each word is tagged with its corresponding part of speech using Charniak model. As assumed earlier, noun and noun phrases are considered as potential features. Therefore these potential features are extracted from each sentence. These are saved in potential features list. The next step is to apply association rule mining to separate out those features which are talked about most. The support is set at 0.02 i.e. 2%. If a feature or feature phrase is talked about in at least 2 review sentences then it will appear in the frequent features. The frequent features obtained after applying Apriori are listed in Table 3.

<<< Insert Table 3 >>>

Along with frequent item sets we also have derived association rules. We have used these association rules for pruning purposes which makes our algorithms better than that of Hu and Liu (2004). The association rules derived from the sample reviews are listed in Table 4.

<<< Insert Table 4 >>>

The features obtained from frequent item set mining may not be meaningful sometime because association rule mining does not take into consideration the sequence of the words in which they appear in the text. In the next step we will perform three types of pruning: compactness pruning, redundancy pruning and miscellaneous pruning. Compactness pruning is used to prune feature phrases that are meaningless. Here we have three two-word feature phrases: "Phone Design", "Battery Life", and "Weight Design". The compactness of these words is given in Table 5. For compactness the distance between the two words is calculated in actual review sentence.

<<< Insert Table 5 >>>

Thus "phone design" and "weight design" will be removed from the features list. In second type of redundancy pruning we prune those single word features that are meaningless when they appear alone. In the current features list, feature like "life" have no meaning when it appears alone. So this will be removed from features list. This is identified with the help of association rules. In the last type of pruning, the features which have same meaning are removed and retained as only one feature. Also, features like "pros" and "cons" are removed. A corpus is used for this purpose. Here we will remove "audio" and make it as "sound". The updated features list after performing pruning is given in Table 6.

<<< Insert Table 6 >>>

In the next step, the opinion words, which are used by reviewers for expressing their opinions about some feature of the product, have been extracted. As explained earlier it is assumed that these opinion words are generally adjectives and are present in the vicinity of feature word or phrase. To execute step nearby adjectives are searched. The features and their opinion words are given in Table 7.

<<< Insert Table 7 >>>

There is no meaning of these opinion words until and unless their semantic orientation is identified. Database is used for this purpose. The seed list of 30 commonly used adjectives with their respective orientation is used. The orientation of all adjectives or opinion words can be found out using this method. In the next step we identified the orientation of full review sentence. The feature-based review-summary is generated in the format given in Figure 1.

The last step is to generate numerical ratings for individual features. The support is set at 5% i.e. frequency of 3. Table 8 presents individual ratings of features.

<<< Insert Table 8 >>>

The results are validated with the help of Crowd Sourcing. We employed 20 human subjects to inspect the results. The results show 93.7% accuracy in summary generation. We have also compared our individual ratings with general public opinion. All the subjects completely agreed with the ratings. These ratings were also shown to the users of the particular product. They agreed with the ratings obtained by our model.

In econometric analysis we have applied GMM model to estimate the parameters. The result shows that there is a significant improvement when we incorporate textual features into the model. The sales and review data was collected from the website for two products: smartphone and digital camera. In Table 9 we present results of 10-fold cross-validation based area under ROC curve (AUC).

<<< Insert Table 9 >>>

## 8. Conclusion and Scope for Future Research

In this study we have focused on the significance of mining opinions from reviews. These opinions are important for both customers and manufacturers. Customers read reviews to make their purchase decision while the manufacturers use these reviews to capture the sentiments of customers all over the web about their products. These reviews are in thousands and thus it becomes time taking process to read all the reviews. Reading few reviews will always facilitate the biased decision rather than informed decision. The first objective of our paper is to summarize these reviews. The results show that we have achieved the accuracy of more than 90% with our approach. The validation is done by several human subjects. The second objective of our paper is to capture the effect of product reviews on product sales. We have collected sales data for two products. We have also performed the prediction of sales for the next week. The result shows a considerable improvement when we incorporate textual variables with numerical review ratings. The product features are very important to customers. Every customer has different preferences of product features. So, a single numerical review rating cannot serve the purpose. To solve this problem, we have generated the individual ratings for all the features. We have shown these ratings to the existing users of the product and they completely agreed with our ratings.

**Scope for Future Research**

In this paper, we have used various Natural Language Processing techniques. The performance of these techniques is really good in all the tasks. But, there is always a scope of improvement. For the future research the accuracy of these algorithms can be improved. In feature extraction, the most crucial part of the research, we have not considered implicit features. Consider an example:

"This phone is too large for my pocket"

In the above example, user is referring to size of the phone. But the word "size" is not explicitly included. These implicit features can be incorporated in the model. For predictive modeling, instead of logistic regression, other machine learning algorithms can be tested for better performance. Finally, a more sophisticated model can be developed to generate review ratings rather than just calculating the proportion of positive sentences.

## References


Archak, N., Ghose, A., and Ipeirotis, P. G. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science*, *57*(8), 1485-1509.

Cao, L., Zhao, Y., Zhang, H., Luo, D., Zhang, C., and Park, E. K. (2010). Flexible frameworks for actionable knowledge discovery. *Knowledge and Data Engineering, IEEE Transactions on*, *22*(9), 1299-1312.

Chen, Y., and Xie, J. (2008). Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management Science*, *54*(3), 477-491.

Das, S. R., and Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*, *53*(9), 1375-1388.

Decker, R., and Trusov, M. (2010). Estimating aggregate consumer preferences from online product reviews. *International Journal of Research in Marketing*,*27*(4), 293-307.

Eliashberg, J., Hui, S. K., and Zhang, Z. J. (2007). From story line to box office: A new approach for green-lighting movie scripts. *Management Science*, *53*(6), 881-893.

Fette, I., Sadeh, N., and Tomasic, A. (2007, May). Learning to detect phishing emails. In *Proceedings of the 16th international conference on World Wide Web*(pp. 649-656). ACM.

Ghose, A., and Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *Knowledge and Data Engineering, IEEE Transactions on*, *23*(10), 1498-1512.

Ghose, A., Ipeirotis, P. G., and Sundararajan, A. (2007, June). Opinion mining using econometrics: A case study on reputation systems. In *Annual Meeting-Association for Computational Linguistics* (Vol. 45, No. 1, p. 416).

Gyongyi, Z., and Garcia-Molina, H. (2005). Web spam taxonomy. In *First international workshop on adversarial information retrieval on the web (AIRWeb 2005)*.

He, J., Zhang, Y., Li, X., and Shi, P. (2012). Learning naive Bayes classifiers from positive and unlabelled examples with uncertainty. *International Journal of Systems Science*, *43*(10), 1805-1825.

Hu, M., and Liu, B. (2004, August). Mining and summarizing customer reviews. In*Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 168-177). ACM.

Li, X., and Hitt, L. M. (2008). Self-selection and information role of online product reviews. *Information Systems Research*, *19*(4), 456-474.

Liu, B., Hu, M., and Cheng, J. (2005, May). Opinion observer: analyzing and comparing opinions on the web. In *Proceedings of the 14th international conference on World Wide Web* (pp. 342-351). ACM.

Mobasher, B., Burke, R., and Sandvig, J. J. (2006, July). Model-based collaborative filtering as a defense against profile injection attacks. In *AAAI*(Vol. 6, p. 1388).

Morinaga, S., Yamanishi, K., Tateishi, K., and Fukushima, T. (2002, July). Mining product reputations on the web. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 341-349). ACM.

Netzer, O., Feldman, R., Goldenberg, J., and Fresko, M. (2012). Mine your own business: Market-structure surveillance through text mining. *Marketing Science*, *31*(3), 521-543.

Öğüt, H., and Onur Taş, B. K. (2012). The influence of internet customer reviews on the online sales and prices in hotel industry. *The Service Industries Journal*,*32*(2), 197-214.

Pang, B., and Lee, L. (2008). Opinion mining and sentiment analysis.*Foundations and trends in information retrieval*, *2*(1-2), 1-135.

Ramezani, M., and Feizi-Derakhshi, M. R. (2014). AUTOMATED TEXT SUMMARIZATION: AN OVERVIEW. *Applied Artificial Intelligence*, *28*(2), 178-215.

Sahami, M., Dumais, S., Heckerman, D., and Horvitz, E. (1998, July). A Bayesian approach to filtering junk e-mail. In Learning for Text Categorization: Papers from the 1998 workshop (Vol. 62, pp. 98-105).

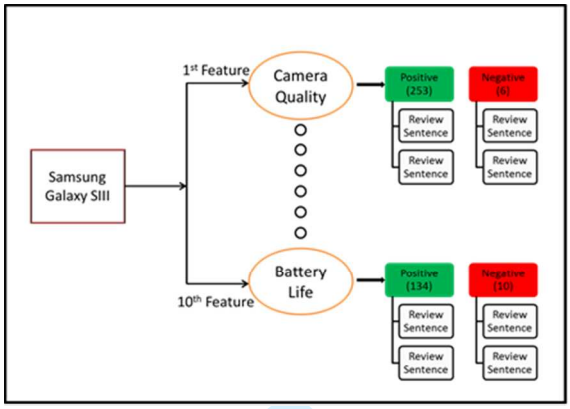


Figure 1. Feature-based summary for smartphone.

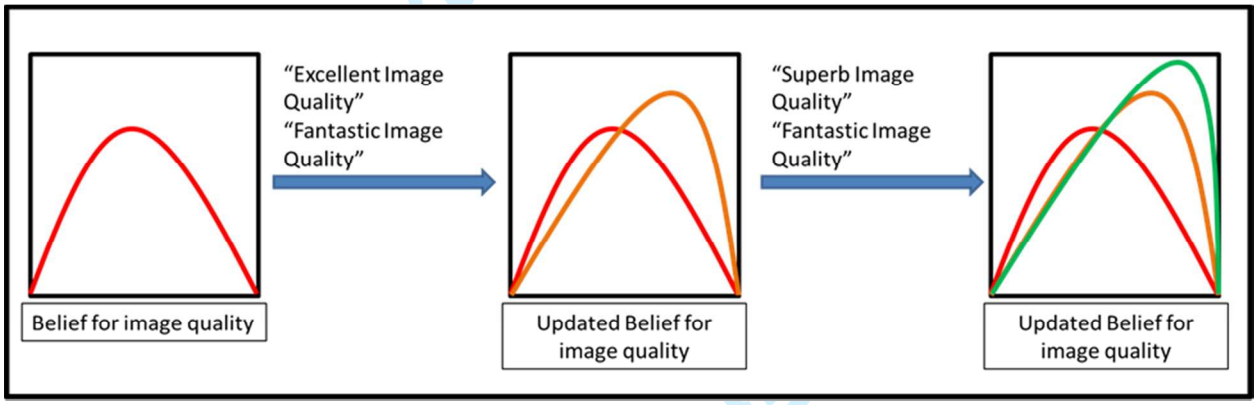


Figure 2. Belief updating from consumer reviews.


Figure 3. General format of reviews.

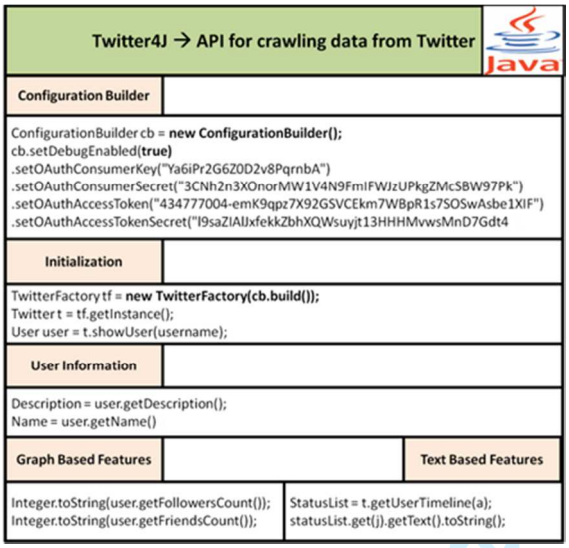


Figure 4. Social media website – Java API.


Figure 5. Spam timeline of social media website feed.

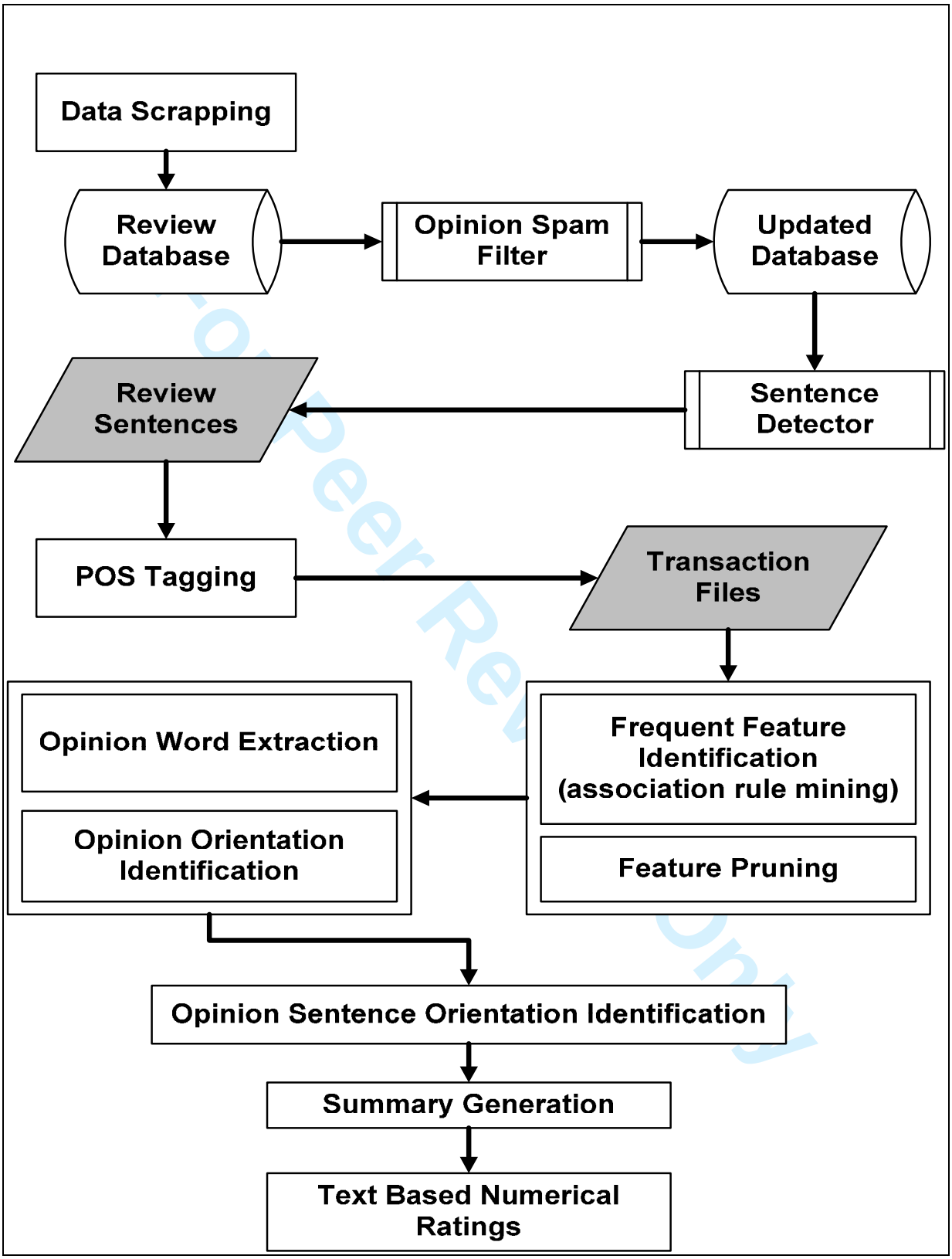


Figure 6. Architectural summary.

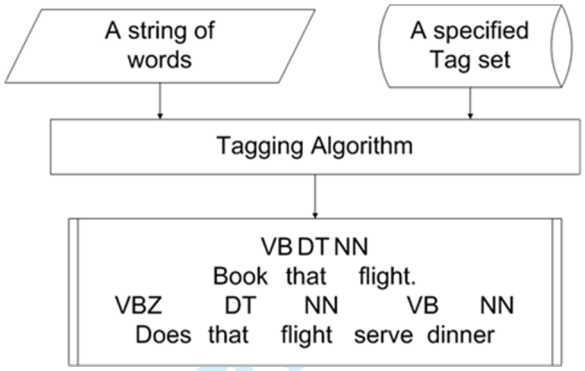


Figure 7. POS Tagging process.

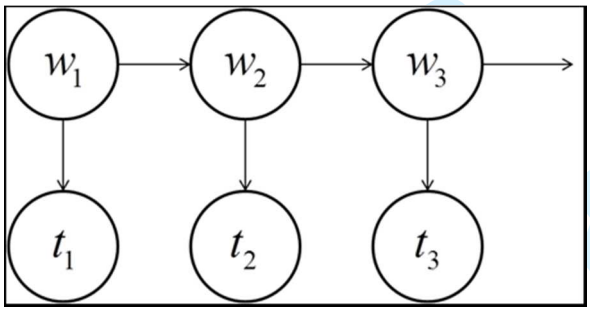


Figure 8. A simple Charniak model.

**Support and Confidence**

• **Support of the rule A → B:**
Denotes the frequency of the rule within all transactions in the database i.e. the probability that a transaction contains both A and B.

• **Confidence of the rule A → B:**
Denotes the percentage of transactions containing A which also contain B, i.e. the probability that a transaction containing A also contains B

Figure 9. Support and Confidence in association rule mining.

```
For each sentence in the review database {

    if (it contains a frequent feature, extract all the adjective
       words as opinion words)
          for each feature in the sentence{

                the nearby adjective is recorded as its  effective
                opinion /* A nearby adjective refers to the adjacent
                adjective that modifies the noun/noun phrase that is
                a frequent feature
                                }
                        }
```

Figure 10. Pseudo code for opinion word extraction.

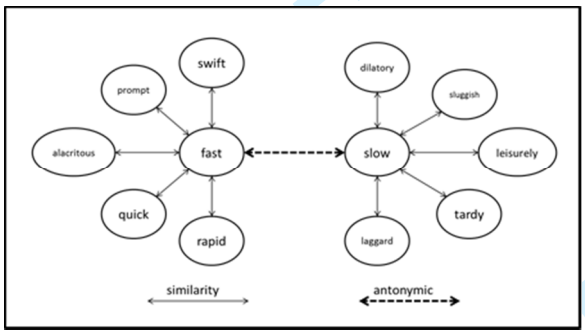


Figure 11. Bipolar structure of Database.

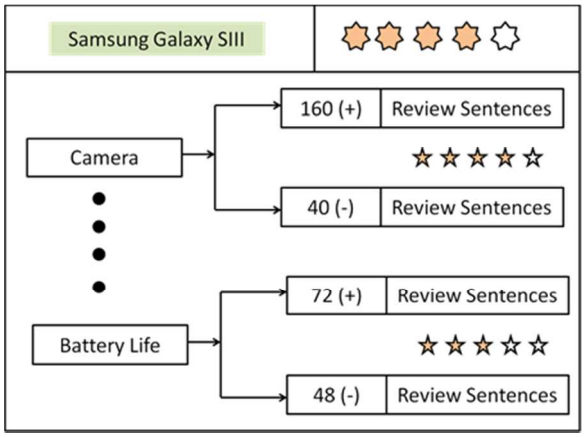


Figure 12. Overall summary of smartphone.

Table 1. The Penn Treebank Part of Speech Tag set.

| Tag | Description | Tag | Description |
|---|---|---|---|
| CC | Coordinating Conjunction | PRP$ | Possessive pronoun |
| CD | Cardinal Number | RB | Adverb |
| DT | Determiner | RBR | Adverb Comparative |
| EX | Existential | RBS | Adverb Superlative |
| FW | Foreign word | RP | Particle |
| IN | Preposition | SYM | Symbol |
| JJ | Adjective | TO | to |
| JJR | Adjective, comparative | UH | Interjection |
| JJS | Adjective, superlative | VB | Verb Base Form |
| LS | List item marker | VBD | Verb Past Tense |
| MD | Modal | VBG | Verb present participle |
| NN | Noun Singular | VBN | Verb past participle |
| NNS | Noun Plural | VBP | Verb, non 3rd person singular present |
| NNP | Proper Noun Singular | VBZ | Verb, 3rd person singular present |
| NNPS | Proper Noun Plural | WDT | Wh-determiner |
| PDT | Pre-determiner | WP | Wh-pronoun |
| POS | Possessive ending | WP$ | Possessive Wh-pronoun |
| PRP | Personal pronoun | WRB | Wh-adverb |

Table 2. A sample dataset for association rule mining.

| ID | Item Sets |
|---|---|
| Review 1 | Word1, Word2, Word3……. |
| Review 2 | Word1, Word4, Word5……. |

Table 3. Frequent features and their frequencies.

| Feature/Feature Phrase | Support | Frequency |
|---|---|---|
| “Hardware” | 0.034 | 2 |
| “Audio” | 0.034 | 2 |
| “Sound” | 0.034 | 2 |
| “Phone” | 0.052 | 3 |
| “Phone Design” | 0.034 | 2 |
| “Android” | 0.034 | 2 |
| “Looks” | 0.052 | 3 |
| “Camera” | 0.103 | 6 |
| “Pros” | 0.034 | 2 |
| “Updates” | 0.052 | 3 |
| “Processor” | 0.069 | 4 |
| “Graphics” | 0.052 | 3 |
| “Display” | 0.069 | 4 |
| “Life” | 0.069 | 4 |
| “Battery Life” | 0.069 | 4 |
| “Battery” | 0.034 | 2 |
| “Weight” | 0.069 | 4 |
| “Weight Design” | 0.034 | 2 |
| “Design” | 0.069 | 4 |
| “Screen” | 0.034 | 2 |
| “Touch” | 0.034 | 2 |

Table 4. Association Rules obtained by Apriori.

| Association Rule | Support | Confidence |
|---|---|---|
| phone -> design | 0.034 | 0.667 |
| design -> phone | 0.034 | 0.500 |
| life -> battery | 0.069 | 1.000 |
| battery -> life | 0.069 | 0.667 |
| weight -> design | 0.034 | 0.500 |
| design -> weight | 0.034 | 0.500 |

Table 5. Compactness of two-word feature phrases.

| Feature Phrase | Frequency | Distance | Result |
|---|---|---|---|
| “Phone Design” | 2 | 9,8 | Not Compact |
| “Battery Life” | 4 | 0,0,0,0 | 100% Compact |
| “Weight Design” | 2 | 6,1 | Not Compact |

Table 6. Updated features list after prunning.

| Feature/Feature Phrase | Support | Frequency |
|---|---|---|
| “Hardware” | 0.034 | 2 |
| “Sound” or “Audio” | 0.069 | 4 |
| “Phone” | 0.052 | 3 |
| “Android” | 0.034 | 2 |
| “Looks” | 0.052 | 3 |
| “Camera” | 0.103 | 6 |
| “Updates” | 0.052 | 3 |
| “Processor” | 0.069 | 4 |
| “Graphics” | 0.052 | 3 |
| “Display” | 0.069 | 4 |
| “Battery Life” | 0.069 | 4 |
| “Battery” | 0.034 | 2 |
| “Weight” | 0.069 | 4 |
| “Design” | 0.069 | 4 |
| “Screen” | 0.034 | 2 |
| “Touch” | 0.034 | 2 |

Table 7. Features and opinion words of their respective sentences.

| Features | Frequency | Opinion Words |
|---|---|---|
| “Hardware” | 2 | { }, {“awesome”} |
| “Sound” | 4 | {“patchy”}, {“bad”},{“bad”}, {“bad”} |
| “Phone” | 3 | { }, {“nice”}, {“best”} |
| “Android” | 2 | {“latest”},{“fast”} |
| “Looks” | 3 | {“good”}, {“excellent”}, {“good”} |
| “Camera” | 6 | {“decent”},{“good”},{“good”},{“average”} {“good”},{“average”} |
| “Updates” | 3 | {“latest”}, {“latest”}, {“better”} |
| “Processor” | 4 | {“fast”},{“good”},{“excellent”} {“awesome”} |
| “Graphics” | 3 | {“superb”}, {“good”}, {“stunning”} |
| “Display” | 4 | {“superb”}, {“good”}, {“good”}, {“amazing”} |
| “Battery Life” | 4 | {“good”},{“average”},{“good”},{“poor”} |
| “Battery” | 2 | { }, {“average”} |
| “Weight” | 4 | {“light”},{“light”},{“light”}, {“light”} |
| “Design” | 4 | {“sleek”},{“beautiful”},{“minimalistic”} {“sleek”} |
| “Screen” | 2 | {“average”},{“fragile”} |
| “Touch” | 2 | { }, {“good”} |

Table 8. Feature-based numerical ratings.

| Features | Frequency | Positive Ratio | Rating (out of 5) |
|---|---|---|---|
| "Sound" or "Audio" | 4 | 0 | 0 |
| "Phone" | 3 | 1 | 5 |
| "Looks" | 3 | 1 | 5 |
| "Camera" | 6 | 0.667 | 3.33 |
| "Updates" | 3 | 1 | 5 |
| "Processor" | 4 | 1 | 5 |
| "Graphics" | 3 | 1 | 5 |
| "Display" | 4 | 1 | 5 |
| "Battery Life" | 4 | 0.5 | 2.5 |
| "Weight" | 4 | 1 | 5 |
| "Design" | 4 | 1 | 5 |

Table 9. Predictive power of empirical model.

| Model | Group | AUC |
|---|---|---|
| No Text | Smartphone | 0.686 |
| Text | Smartphone | 0.798 |
| No Text | Digital Camera | 0.713 |
| Text | Digital Camera | 0.831 |

# Notes on Contributors

***Akshay Kangale*** is currently pursuing his dual degree in Manufacturing Science & Engineering/ Industrial Engineering & Management from Indian Institute of Technology Kharagpur, India. His research interest lies in the field of operations research and supply chain.

***Professor S Krishna Kumar*** is faculty member in the department of Industrial & Systems Engineering, Indian Institute of Technology Kharagpur, India. He works in the area of operations research, supply chain and logistics, non-linear programming and game theory.

***Mohd Arshad Naeem*** is presently working in "American Express" and has completed his dual degree in Industrial Engineering/Industrial Engineering & Management from Indian Institute of Technology Kharagpur, India.

***Professor Mark Williams*** is leader of the Product Evaluation Technologies (PET) and Metrology Research groups in WMG at the University of Warwick. In 1990, he served his time as a Technician Apprentice at GEC Alsthom and studied Mechanical Engineering at Nottingham Trent University. He then received a PhD in Engineering from the University of Manchester Institute of Science and Technology (UMIST) in 1998. Mark is a also a Chartered Engineer (CEng) and Fellow of the Institution of Mechanical Engineers (FIMechE) and has worked within a wide range of industries as a Dimensional Control consultant and was employed at Jaguar Land Rover up until 2003 when he joined Warwick Manufacturing Group (WMG). Whilst at WMG, Prof Williams has managed a diverse portfolio of research projects funded by EPSRC, the Technology Strategy Board (TSB), Catapult and EU within the Automotive, Aerospace, Healthcare and Defense industries. This research has led to the publication of over 50 academic papers published in peer reviewed journals and presented at international conferences.

***Professor M.K. Tiwari*** is the head of the department of Industrial & Systems Engineering, Indian Institute of Technology Kharagpur, India. He has 19 years of teaching and research experience at different levels. He works in the area of evolutionary computing, applications, modeling and simulation of manufacturing system, supply chain management, planning and scheduling of automated manufacturing system, production planning and control, etc. He is listed among top 20 Most Productive Authors in the area of production and operations management as reported in the last 50 years and rated second among many researchers working in Logistics and Supply Chain Management in India. He has published around 195 articles in leading international journals and serves on the editorial board of around seven international journals. He is serving as Associate editors of IEEE System, Man and Cybernetics- System,(SMC-S), International Journal of Production Research(Taylor and Francis), Computers and Industrial Engineering (Elsevier), International Journal of System Science (Taylor and Francis) and Journal of Intelligent Manufacturing System (Springer) and Neuro-computing(Elsevier).